10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Voice Conversion without Explicit Separation of Source and Filter Components Based on Non-negative Matrix Factorization

Hitoshi Suda[1], *Daisuke Saito*[1], *Nobuaki Minematsu*[1]

[1]Graduate School of Engineering, The University of Tokyo, Japan

{hitoshi,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract

This paper introduces a new voice conversion (VC) technique which performs spectrogram-to-spectrogram conversion. Conventional studies on VC focus on spectral envelopes, which represent vocal tract information. While vocoders have enabled light-weight and high-quality synthesis from the features, flexibility and quality is still limited by parameterization. To overcome the limitation, this paper aims to model and convert spectrograms themselves. In general, spectrograms are too complicated to be modeled because they contain not only spectral envelopes but also source structures. This paper adopts source-filter non-negative matrix factorization (SF-NMF) as a conversion model of spectrograms. SF-NMF is an extended model of non-negative matrix factorization (NMF), and models source and filter components jointly without explicit separation. The proposed method generates waveforms by reconstructing phase information from amplitude spectrograms. Since SF-NMF requests log-frequency spectrograms, the method utilizes scalograms, which are obtained by continuous wavelet transform (CWT). Experimental results showed the proposed method achieved spectrogram-to-spectrogram speaker conversion.

Index Terms: voice conversion, source-filter non-negative matrix factorization, continuous wavelet transform, phase reconstruction

1. Introduction

Voice conversion (VC), or speaker conversion, is a technique to make an utterance sound like that of a different speaker while keeping its linguistic contents [1]. VC systems generally consist of following 3 steps: extraction of acoustic features that contains speaker information, conversion of the extracted features, and waveform synthesis from the converted features. As acoustic features, spectral envelopes or mel-cepstral coefficients are widely used. Gaussian mixture models (GMMs) and non-negative matrix factorization (NMF) are conventionally used to model non-linear conversion of these features [2, 3]. Recent advances on studies on neural networks (NNs) enable more complex and sophisticated modeling of conversion [4, 5].

To synthesize waveforms from converted features, analysis-synthesis systems such as STRAIGHT [6] and WORLD [7] are often used. While these vocoders enable light-weight and high-quality synthesis, parameterized representation of speech can degrade synthesized utterances. To overcome this defect, neural synthesis methods such as WaveNet vocoder have been proposed [8]. However, even if these sophisticated synthesizers are utilized, flexibility in representation of speech is still restricted by parameterization.

As an alternate synthesis method, waveform reconstruction from amplitude spectrograms has been studied over years. Griffin-Lim method infers phase information from given spectrograms by minimizing mean square errors, and generates as consistent waveforms as possible [9]. An NN-based phase reconstruction technique has also been proposed by utilizing von Mises distribution and evaluating group delays [10]. VC systems are expected to be more flexible if spectrograms themselves are converted and these techniques are applied. While spectrograms are versatile representation of waveforms, they are so complicated that conventional conversion models cannot handle them.

Source-filter NMF (SF-NMF) is an extended model of NMF which can handle spectrograms of mixed sounds such as polyphonic music signals [11]. This paper manipulates spectrograms using SF-NMF, and accomplishes VC in a similar way to NMF-based VC. Since SF-NMF can handle polyphonic signals, it is so complex for single speaker's utterance and computationally expensive. To solve this problem, this paper proposes simplified SF-NMF which can handle monophonic audio signals.

The remainder of this paper is organized as follows. Section 2 describes conventional NMF-based VC systems and proposed SF-NMF-based frameworks. Section 3 details analysis and synthesis techniques of spectrograms using continuous wavelet transform. Section 4 describes subjective experiments conducted to evaluate the quality of the proposed framework. Section 5 concludes the paper.

2. SF-NMF-based VC framework

2.1. Non-negative matrix factorization

NMF is a group of algorithm to decompose a non-negative matrix into multiplication of two non-negative matrices [12]. Let $\boldsymbol{Y} \in \mathbb{R}^{K\times T}$ be an input matrix, and $\boldsymbol{H} \in \mathbb{R}^{K\times N}$ and $\boldsymbol{U} \in \mathbb{R}^{N\times T}$ be decomposed matrices. The main problem of NMF is to obtain $\boldsymbol{H}$ and $\boldsymbol{U}$ that satisfy

$$\boldsymbol{Y} \approx \boldsymbol{H}\boldsymbol{U}, \tag{1}$$

where all the elements of $\boldsymbol{Y}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are non-negative. Supposing that $\boldsymbol{Y} = [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{T-1}]$ is a spectrogram, (1) can be written as

$$\boldsymbol{y}_t = \sum_{n=0}^{N-1} u_{n,t}\boldsymbol{h}_n, \tag{2}$$

that is, a spectrum $\boldsymbol{y}_t$ is represented as weighted summation of N spectral templates $\boldsymbol{h}_0, \ldots, \boldsymbol{h}_{N-1}$. Each $\boldsymbol{h}_n$ is regarded as a template and $u_{n,t}$ indicates how greatly the template appears at time t. Hence, $\boldsymbol{H}$ and $\boldsymbol{U}$ are called *dictionary* and *activation*, respectively.

$\boldsymbol{H}$ and $\boldsymbol{U}$ can be optimized by minimizing difference between $\boldsymbol{Y}$ and $\boldsymbol{H}\boldsymbol{U}$, or

$$\mathcal{D}(\boldsymbol{Y}|\boldsymbol{H}\boldsymbol{U}) \to \min., \tag{3}$$

10.21437/SSW.2019-13

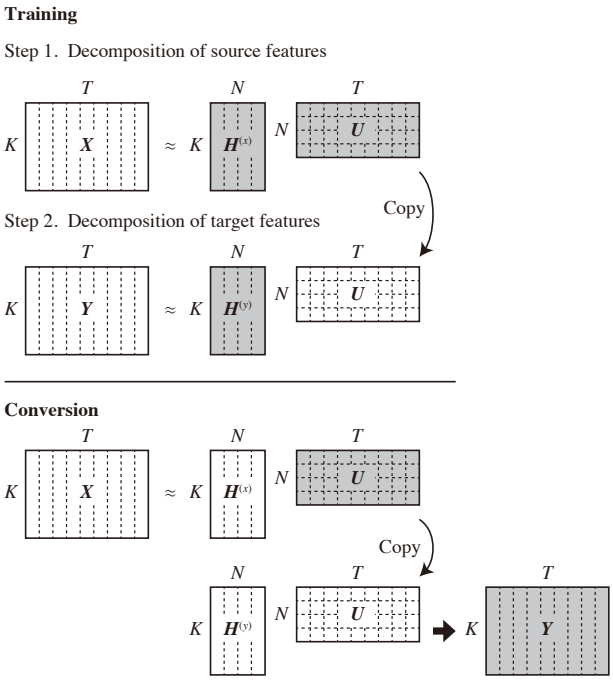


Figure 1: *Overview of conventional NMF-based VC frameworks [3]. Gray-colored matrices are estimated or calculated in each process.*

where $\mathcal{D}$ denotes a divergence function such as Euclidean distance or generalized Kullback-Leibler (KL) divergence. For decomposition of amplitude spectrograms, generalized KL divergence is adopted as $\mathcal{D}$, which is defined as

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{A}|\boldsymbol{B}) = \sum_{k,t}\left(A_{kt}\log\frac{A_{kt}}{B_{kt}} - A_{kt} + B_{kt}\right). \quad (4)$$

This optimization problem cannot be solved analytically, and thus an iterative algorithm is provided based on auxiliary function method [12].

2.2. NMF-based VC framework

Figure 1 summaries the basic NMF-based VC framework described in this section.

Let $\boldsymbol{X} = [\boldsymbol{x}_0, \dots, \boldsymbol{x}_{T-1}]$ and $\boldsymbol{Y} = [\boldsymbol{y}_0, \dots, \boldsymbol{y}_{T-1}]$ be time-aligned sequences of acoustic features extracted from source and target speakers' utterances respectively which have the same linguistic contents. In NMF-based VC frameworks, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are decomposed into shared activation $\boldsymbol{U}$ and speaker-dependent dictionaries $\boldsymbol{H}^{(x)}$ and $\boldsymbol{H}^{(y)}$, that is,

$$\boldsymbol{x}_t \approx \sum_{n=0}^{N-1} u_{n,t}\boldsymbol{h}_n^{(x)} \quad \text{and} \quad \boldsymbol{y}_t \approx \sum_{n=0}^{N-1} u_{n,t}\boldsymbol{h}_n^{(y)}. \quad (5)$$

Once these approximations are fulfilled, the correspondence of $\boldsymbol{h}_n^{(x)}$ and $\boldsymbol{h}_n^{(y)}$ is taken on each index n. Therefore, $\boldsymbol{H}$ and $\boldsymbol{U}$ can be considered to contain speaker and linguistic information, respectively.

In conversion process, activation $\boldsymbol{U}$ is estimated based on an input feature sequence $\boldsymbol{X}$ and the source speaker's dictionary $\boldsymbol{H}^{(x)}$, and then a feature sequence of the target speaker $\boldsymbol{Y}$ is calculated by $\boldsymbol{Y} = \boldsymbol{H}^{(y)}\boldsymbol{U}$.

2.3. SF-NMF

Since NMF represents input matrices as weighted summation of templates, it is difficult to model spectrograms with source structures by NMF. To get over this difficulty, this paper utilizes SF-NMF, which is an extended model of NMF and can deal with spectrograms even with source structures [11]. SF-NMF decomposes an input matrix $\boldsymbol{Y} \in \mathbb{R}^{K\times T}$ by following equation:

$$y_t(k) \approx \sum_{n,m,z} u_{n,m,z,t}e_n(k-z)h_m(k), \quad (6)$$

where $\boldsymbol{E} \in \mathbb{R}^{K\times N}$ and $\boldsymbol{H} \in \mathbb{R}^{K\times M}$ are dictionaries for source and filter components respectively, and $\boldsymbol{U} \in \mathbb{R}^{N\times M\times Z\times T}$ denotes their activation. SF-NMF can handle multiple sources with different spectral envelopes, and therefore this model is so complicated for single speaker's utterances. To solve this problem, this paper proposes a simplified SF-NMF defined by

$$y_t(k) \approx \left(\sum_z u_{z,t}^{(e)}e(k-z)\right)\left(\sum_m u_{m,t}^{(h)}h_m(k)\right) + \sum_i u_{i,t}^{(a)}a_i(k). \quad (7)$$

Here, $\boldsymbol{e} \in \mathbb{R}^K$, $\boldsymbol{H} \in \mathbb{R}^{K\times M}$, and $\boldsymbol{A} \in \mathbb{R}^{K\times I}$ denote dictionaries for source, filter, and aperiodic components, respectively, and $\boldsymbol{U}^{(e)} \in \mathbb{R}^{Z\times T}$, $\boldsymbol{U}^{(h)} \in \mathbb{R}^{M\times T}$, and $\boldsymbol{U}^{(a)} \in \mathbb{R}^{I\times T}$ are activation of the corresponding dictionaries. Equation (7) models spectra as summation of harmonic and aperiodic components, that is,

$$y_t(k) \approx \alpha_t(k)\beta_t(k) + \gamma_t(k), \quad (8)$$

where $\alpha_t(k) = \sum_z u_{z,t}^{(e)}e(k-z)$, $\beta_t(k) = \sum_m u_{m,t}^{(h)}h_m(k)$, and $\gamma_t(k) = \sum_i u_{i,t}^{(a)}a_i(k)$ represent source, filter, and aperiodic components, respectively. Figure 2 shows an overview of this model. See Appendix A for the update functions and their derivation.

SF-NMF can be utilized for conversion model in a similar way to NMF. Note that activation $\boldsymbol{U}^{(e)}$ of target utterances cannot be identical to that of source ones in training step even if the linguistic contents are the same.

3. Scalogram and its phase reconstruction

Since SF-NMF models source structures as weighted summation of shifted templates, decomposed matrices should be log-frequency spectrograms. This section describes continuous wavelet transform, which can obtain desired spectrograms without interpolation, and a phase reconstruction technique from obtained spectrograms.

3.1. Continuous wavelet transform

Continuous wavelet transform (CWT)[1] is an alternative method for time-frequency analysis, which can obtain spectrograms with arbitrary frequency resolution and basis function [13]. Basis functions of CWT are called *wavelet*s, and obtained spectrograms are called *scalogram*s. A scalogram $\boldsymbol{s}$ with frequency

[1] In this paper, CWT is performed in discrete time, and this can be called discrete wavelet transform (DWT). However, some tools call multiresolution analysis DWT. To avoid this confusion, this paper calls it CWT even for discrete-time-series data.

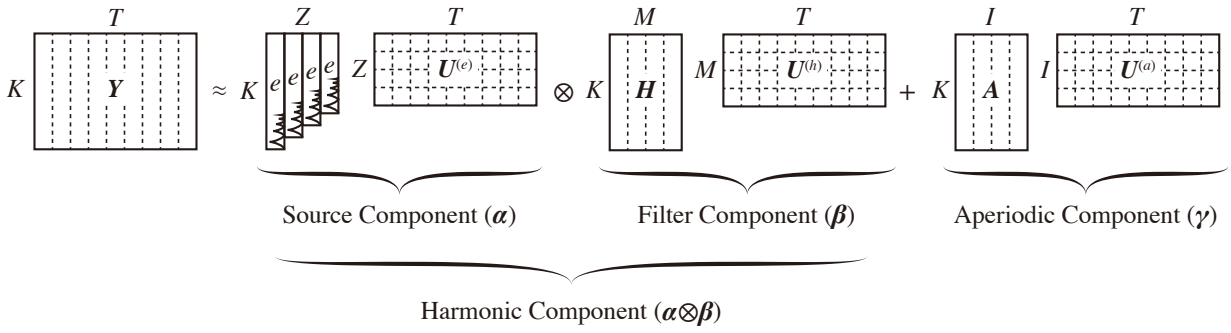


Figure 2: *Conceptual diagram of simplified SF-NMF proposed in this paper. This model decomposes whole spectrogram into source, filter, and aperiodic components internally.*

bins $f_0, \ldots, f_{L-1}$ based on mother wavelet $\psi(t)$ can be obtained by following equations:

$$\boldsymbol{s}_l = \boldsymbol{W}_l \boldsymbol{x} \tag{9}$$

$$\boldsymbol{s} = \left[\boldsymbol{s}_0^\top, \ldots, \boldsymbol{s}_{L-1}^\top\right]^\top = \left[\boldsymbol{W}_0^\top, \ldots, \boldsymbol{W}_{L-1}^\top\right]^\top \boldsymbol{x} \tag{10}$$

$$= \boldsymbol{W}\boldsymbol{x}, \tag{11}$$

$$\boldsymbol{W}_l = \begin{bmatrix} \psi_{l,0} & \psi_{l,1} & \cdots & \psi_{l,T-1} \\ \psi_{l,T-1} & \psi_{l,0} & \cdots & \psi_{l,T-2} \\ \vdots & & \ddots & \vdots \\ \psi_{l,1} & \psi_{l,2} & \cdots & \psi_{l,0} \end{bmatrix}, \tag{12}$$

where $l \in [0, L)$ and $t \in [0, T)$ denote indices of frequency bins and discrete time respectively, $\psi_{l,t} = \psi(t\Delta/f_l)/\sqrt{f_l}$ is a scaled wavelet whose center frequency is f_l, and Δ denotes sampling period.

Since CWT is convolution of waveforms and wavelets, scalograms can be obtained easily in frequency domain by following equations:

$$\hat{\boldsymbol{s}}_l = \hat{\boldsymbol{W}}_l \hat{\boldsymbol{x}} \tag{13}$$

$$\hat{\boldsymbol{s}} = \left[\hat{\boldsymbol{s}}_0^\top, \ldots, \hat{\boldsymbol{s}}_{L-1}^\top\right]^\top = \left[\hat{\boldsymbol{W}}_0^\top, \ldots, \hat{\boldsymbol{W}}_{L-1}^\top\right]^\top \hat{\boldsymbol{x}} \tag{14}$$

$$= \hat{\boldsymbol{W}}\hat{\boldsymbol{x}}, \tag{15}$$

$$\hat{\boldsymbol{W}}_l = \boldsymbol{F}_T \boldsymbol{W}_l \boldsymbol{F}_T^{-1} = \mathrm{diag}\left[\hat{\psi}_{l,0}, \ldots, \hat{\psi}_{l,T-1}\right], \tag{16}$$

$$\hat{\boldsymbol{s}} = \boldsymbol{F}_T \boldsymbol{s}, \quad \hat{\boldsymbol{x}} = \boldsymbol{F}_T \boldsymbol{x}, \quad \hat{\boldsymbol{\psi}}_l = \boldsymbol{F}_T \boldsymbol{\psi}_l. \tag{17}$$

Here, $\boldsymbol{F}_T \in \mathbb{C}^{T\times T}$ denotes discrete Fourier transform matrix, which is defined by

$$(\boldsymbol{F}_T)_{k,t} = e^{-i\frac{2\pi kt}{T}}. \tag{18}$$

3.2. Inverse CWT and phase reconstruction

Inverse CWT is regarded as an optimization problem that finds most consistent waveforms $\tilde{\boldsymbol{x}}$ with a given scalogram $\boldsymbol{s}$, that is,

$$\tilde{\boldsymbol{x}} = \arg\min_{\boldsymbol{\xi}} \|\boldsymbol{s} - \boldsymbol{W}\boldsymbol{\xi}\|. \tag{19}$$

Based on a minimum mean square error criterion, $\tilde{\boldsymbol{x}}$ can be derived by following equations:

$$\tilde{\boldsymbol{x}} = \boldsymbol{W}^+ \boldsymbol{s}, \quad \boldsymbol{W}^+ = \left(\boldsymbol{W}^\dagger \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\dagger, \tag{20}$$

where $\boldsymbol{W}^+$ is the Moore-Penrose-inverse of $\boldsymbol{W}$, and † denotes Hermitian conjugate [14]. In frequency domain, inverse CWT can be written down as

$$\hat{\tilde{x}}_t = \frac{\sum_l \hat{\psi}^*_{l,t} \hat{s}_{l,t}}{\sum_l \hat{\psi}^*_{l,t} \hat{\psi}_{l,t}}, \tag{21}$$

where * denotes complex conjugate.

Since SF-NMF cannot handle scalograms with phase information, phase reconstruction must be performed for waveform generation. In the same way of Griffin-Lim algorithm for short-time Fourier transform spectrogram [9], phase reconstruction of scalograms is defined as an optimization problem, that is,

$$\tilde{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \left\| \boldsymbol{a}e^{i\boldsymbol{\phi}} - \boldsymbol{W}\boldsymbol{W}^+ \boldsymbol{a}e^{i\boldsymbol{\phi}} \right\|, \tag{22}$$

where $\boldsymbol{a} \in \mathbb{R}^{LT}$ denotes a given amplitude scalogram, and $\tilde{\boldsymbol{\phi}} \in [-\pi, \pi)^{LT}$ is estimated phase information [15]. Using auxiliary function method, $\tilde{\boldsymbol{\phi}}$ can be optimized by iteration of following equations:

$$\tilde{\boldsymbol{s}} \leftarrow \boldsymbol{W}\boldsymbol{W}^+ \boldsymbol{a}e^{i\tilde{\boldsymbol{\phi}}}, \quad \tilde{\boldsymbol{\phi}} \leftarrow \angle\tilde{\boldsymbol{s}}. \tag{23}$$

3.3. Fast CWT and phase reconstruction

Since wavelets have limited bandwidth, generic CWT gives redundant scalograms. By limiting bandwidth of wavelets approximately, fast CWT can be performed by band limitation and phase circulation [15]. Supposing a wavelet $\hat{\boldsymbol{\psi}}_l$ can be clipped with $t \in [B_l, B_l + D)$, a band-limited scalogram $\boldsymbol{s'} \in \mathbb{C}^{LD}$ is defined as

$$\boldsymbol{s}' = \left[\boldsymbol{s}_0'^\top, \ldots, \boldsymbol{s}_{L-1}'^\top\right]^\top, \tag{24}$$

$$\hat{\boldsymbol{s}}'_l = \boldsymbol{F}_D \boldsymbol{s}'_l = \hat{\boldsymbol{W}}'_l \hat{\boldsymbol{x}} = \frac{1}{q} \boldsymbol{C}_l \boldsymbol{L}_l \hat{\boldsymbol{W}}_l \hat{\boldsymbol{x}}, \tag{25}$$

$$\boldsymbol{C}_l = \begin{bmatrix} \boldsymbol{0}_{C_l \times (D-C_l)} & \boldsymbol{I}_{C_l} \\ \boldsymbol{I}_{D-C_l} & \boldsymbol{0}_{(D-C_l)\times C_l} \end{bmatrix}, \tag{26}$$

$$\boldsymbol{L}_l = \begin{bmatrix} \boldsymbol{0}_{D\times B} & \boldsymbol{I}_D & \boldsymbol{0}_{D\times(T-B-D)} \end{bmatrix}, \tag{27}$$

where $C_l = B_l - D\lceil B_l/D - 1\rceil$. Obtained $\boldsymbol{s}'$ has a property that

$$\boldsymbol{s}'_{l,d} \sim \boldsymbol{s}_{l,(T/D)d}. \tag{28}$$

In the same way as (20), inverse transform can be derived as follows:

$$\hat{\tilde{\boldsymbol{x}}} = \hat{\boldsymbol{W}}_l'^+ \boldsymbol{s}', \tag{29}$$

or

$$\hat{\tilde{x}}_t = \frac{\sum_l \hat{\psi}'^*_{l,t} \hat{s}'_{l,r_{l,t}}}{\sum_l \hat{\psi}'^*_{l,t} \hat{\psi}'_{l,t}}, \tag{30}$$

where

$$\hat{\psi}'_{l,t} = \begin{cases} \hat{\psi}_{l,t} & (B_l \leq t < B_l + D) \\ 0 & (\text{otherwise}) \end{cases}, \tag{31}$$

$$r_{l,B_l+\tau} = \begin{cases} C_l + \tau & (0 \leq \tau < D - C_l) \\ C_l - D + \tau & (D - C_l \leq \tau < D) \end{cases}. \tag{32}$$

Also, phase reconstruction for fast CWT can be performed in the same way as (23).

4. Experiment of VC systems

This section details subjective experiments conducted to evaluate the quality of SF-NMF-based VC compared with a conventional NMF-based VC system.

4.1. Experimental setups

The prepared data were the speech data of 2 Japanese speakers that uttered ATR Japanese phonetically balanced sentence sets [16]. In the experiments, the subset A (50 sentences) was used for training, and the subset J (53 sentences) for evaluation. The source speaker was a male speaker, and the target was a female one. The sampling frequency was 16 kHz. Time alignment between source and target features was performed based on affine-DTW [17].

In SF-NMF-based system, the number of dictionaries was 1 for source (e), 200 for filter (h), and 5 for aperiodicity (a). The maximum number of shifts of source dictionary was set to 96, that is, 2 octaves. Each dictionary of source was initialized by a scalogram of an impulse sequence whose frequency is a half of the mean of the fundamental frequencies. The activation of that was initialized by fundamental frequencies obtained by WORLD analysis [7]. The mother wavelet was log-normal wavelet [18], which is defined by

$$\hat{\psi}(\omega) = \begin{cases} \exp\left(-\frac{(\log \omega)^2}{4\sigma^2}\right) & (\omega > 0) \\ 0 & (\omega \leq 0) \end{cases}, \tag{33}$$

and σ was set to $\sqrt{2}/100$. The number of frequency bins was 349, and the range was from 50 Hz to $50 \times 2^{7.25} \approx 7611$ Hz (48 bins per oct.).

In NMF-based system, the number of dictionaries was set to 200. WORLD [7] (D4C Edition [19]) was used for analysis and synthesis. Fundamental frequencies were converted linearly based on the means.

Subjective listening tests were executed to evaluate naturalness and speaker similarity of converted utterances. In naturalness tests, 29 subjects answered how natural the utterances are. In speaker similarity tests, 28 subjects answered which speaker is more similar to that of converted utterances, source or target. In both tests, each subject evaluated the utterances in 5-point scale. The tests were executed via our crowdsourcing system. Each participant answered 10 questions and earned approximately 0.45 dollars for his/her participation.

4.2. Results

Figure 3 shows the results of the tests about speaker similarity. The results show the SF-NMF-based system converted speaker identity at the comparable quality to the NMF-based system, and indicate that the concept of NMF-based VC in the proposed framework did not miss.

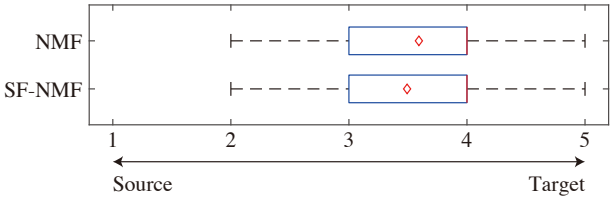


Figure 3: *Results of subjective experiments about speaker similarity of conversion utterances. ◇ denotes the mean of the scores.*

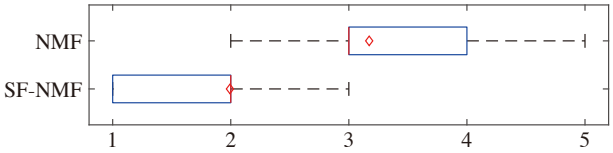


Figure 4: *Results of subjective experiments about naturalness of conversion utterances. ◇ denotes the mean of the scores.*

Figure 4 shows the results of comparison about naturalness. The SF-NMF-based framework generated waveforms of the lower quality compared with NMF-based VC. On the other hand, the results also show the quality of utterances were not so low that all participants answered completely unnatural.

4.3. Discussions

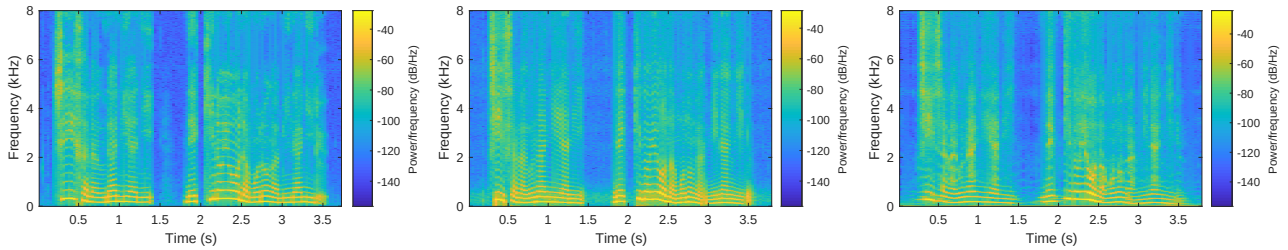
Although the proposed framework achieved speaker conversion, the results show the quality of the generated utterances sounded lower than that of the conventional method.

Figure 5 shows spectrograms of target and generated waveforms. On the frequencies below 2 kHz, source structures can be observed in all the spectrograms. However, on the frequencies over 2 kHz, the spectrogram obtained by SF-NMF has fuzzier source structures than the others.

Figure 6 shows the trained source dictionaries in the proposed system. Similarly to spectrograms, the source dictionaries are also fuzzy on over 2 kHz. Since variances of log-normal wavelets are constant in log-frequency domain (see (33)), the shapes of wavelets are duller on higher frequencies in linear-frequency domain. In other words, scalograms obtained based on the wavelets do not have sufficient expressiveness for speech synthesis. This is because scalograms are fuzzy on over 2 kHz, and this caused noisy sounds. Therefore, sharper wavelets can solve this fuzziness problem.

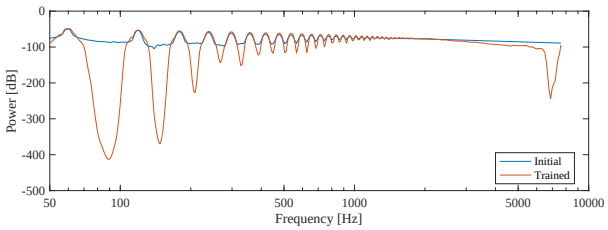
5. Conclusions

This paper has proposed a new VC framework based on SF-NMF, which provides spectrogram-to-spectrogram conversion. Since the model can handle spectrograms without explicit separation of source and filter components, SF-NMF-based VC can also convert the source structures of input waveforms. The experimental results have shown that speaker conversion was achieved while the quality of synthesized speech was not comparable to that of vocoders. The proposed framework was not able to model detail structures on the high frequencies. This is because the used wavelets lack of sharpness on the higher fre-

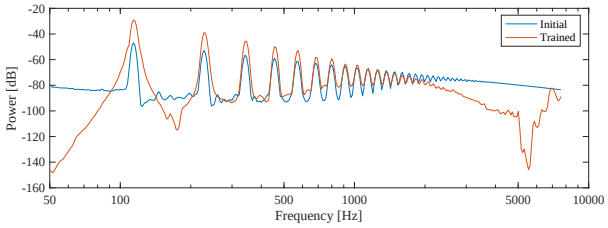


(a) *Target spectrogram (natural utterance)* (b) *NMF* (c) *SF-NMF*

Figure 5: *Comparison of spectrograms of the target and generated waveforms. Compared to (b) NMF, (c) SF-NMF obtains fuzzier spectrograms on higher frequencies. The sentence is j01 of ATR phonetically balanced sentence sets.*



(a) *Dictionaries of the source speaker.*



(b) *Dictionaries of the target speaker.*

Figure 6: *Comparison of dictionaries for source (excitation). In each figure, blue denotes initial dictionary, and orange denotes trained one. All the peaks have the same width in log-frequency domain.*

quencies, and this is an explanation of degradation of generated samples.

For future works, the effectiveness of wavelets should be investigated. The experimental results in the paper indicate appropriate wavelets can be different for analysis and synthesis. Also, techniques for phase reconstruction should be investigated. In this paper, phase reconstruction based on consistency was performed. However, the quality of consistent waveforms is not always perceptually high. To overcome this defect, some deep-neural-network-based approaches have been proposed [10]. By utilizing these techniques, more natural sounds will be synthesized.

Since the proposed model has flexibility in modeling of source structures, the model has the ability to handle abnormal utterances such as gravelly or creaky voices. Hence, the application to analysis and conversion of such the utterances is expected.

6. Acknowledgements

This research and development work was supported by the Ministry of Internal Affairs and Communications.

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan*, vol. 11, no. 2, pp. 71–76, 1990.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[3] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *2012 IEEE Spoken Language Technology Workshop*, 2012, pp. 313–317.

[4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3893–3896.

[5] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH 2017*, 2017, pp. 1283–1287.

[6] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[7] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[8] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *INTERSPEECH 2017*, 2017, pp. 1118–1122.

[9] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[10] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," in *16th International Workshop on Acoustic Signal Enhancement*, 2018, pp. 286–290.

[11] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop 18*, 2006.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 556–562.

[13] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998.

[14] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," vol. 41, no. 12, pp. 3549–3554, 1993.

[15] T. Nakamura and H. Kameoka, "Fast signal reconstruction from magnitude spectrogram of continuous wavelet transform based on spectrogram consistensy," in *17th International Conference on Digital Audio Effects*, 2014, pp. 1–7.

[16] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[17] H. Suda, G. Kotani, S. Takamichi, and D. Saito, "A revisit to feature handling for high-quality voice conversion based on Gaussian mixture model," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, pp. 816–822.

[18] H. Kameoka, "Statistical approach to multipitch analysis," Ph.D. Thesis, The University of Tokyo, 2007.

[19] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

A. Update equations for simplified SF-NMF and their derivation

This appendix describes the update equations of the proposed model and derivation of them.

First, let us define a generalized form of the proposed model by

$$y_t(k) \approx \sum_{j=0}^{J-1} \alpha_{j,t}(k)\beta_{j,t}(k), \tag{34}$$

$$\alpha_{j,t}(k) = \sum_{z=0}^{Z_j-1} u^{(e)}_{j,z,t} e_j(k-z), \tag{35}$$

$$\beta_{j,t}(k) = \sum_{m=0}^{M_j-1} u^{(h)}_{j,m,t} h_{j,m}(k). \tag{36}$$

By the auxiliary function method, the following update equations of this model can be derived:

$$e_j(k) \leftarrow e_j(k) \frac{\sum_{t,z} \frac{y_t(k+z)}{x_t(k+z)} \beta_{j,t}(k+z) u^{(e)}_{j,z,t}}{\sum_{t,z} \beta_{j,t}(k+z) u^{(e)}_{j,z,t}}, \tag{37}$$

$$u^{(e)}_{j,z,t} \leftarrow u^{(e)}_{j,z,t} \frac{\sum_k \frac{y_t(k)}{x_t(k)} \beta_{j,t}(k) e_j(k-z)}{\sum_k \beta_{j,t}(k) e_j(k-z)}, \tag{38}$$

$$h_{j,m}(k) \leftarrow h_{j,m}(k) \frac{\sum_t \frac{y_t(k)}{x_t(k)} \alpha_{j,t}(k) u^{(h)}_{j,m,t}}{\sum_t \alpha_{j,t}(k) u^{(h)}_{j,m,t}}, \tag{39}$$

$$u^{(h)}_{j,m,t} \leftarrow u^{(h)}_{j,m,t} \frac{\sum_k \frac{y_t(k)}{x_t(k)} \alpha_{j,t}(k) h_{j,m}(k)}{\sum_k \alpha_{j,t}(k) h_{j,m}(k)}, \tag{40}$$

where

$$x_t(k) = \sum_{j=0}^{J-1} \alpha_{j,t}(k)\beta_{j,t}(k). \tag{41}$$

The proposed model is a specilialized form of (34) with following constraints:

$$J = 2,\ \boldsymbol{e}_1 = \mathbf{1},\ \boldsymbol{u}^{(e)}_{z=1} = \mathbf{1},\ Z_1 = 0. \tag{42}$$

By applying these constraints, the desired update equations for (7) are derived as follows:

$$e(k) \leftarrow e(k) \frac{\sum_{z,t} r_t(k+z)\beta_t(k+z) u^{(e)}_{z,t}}{\sum_{z,t} \beta_t(k+z) u^{(e)}_{z,t}}, \tag{43}$$

$$u^{(e)}_{z,t} \leftarrow u^{(e)}_{z,t} \frac{\sum_k r_t(k)\beta_t(k) e(k-z)}{\sum_k \beta_t(k) e(k-z)}, \tag{44}$$

$$h_m(k) \leftarrow h_m(k) \frac{\sum_t r_t(k)\alpha_t(k) u^{(h)}_{m,t}}{\sum_t \alpha_t(k) u^{(h)}_{m,t}}, \tag{45}$$

$$u^{(h)}_{m,t} \leftarrow u^{(h)}_{m,t} \frac{\sum_k r_t(k)\alpha_t(k) h_m(k)}{\sum_k \alpha_t(k) h_m(k)}, \tag{46}$$

$$a_i(k) \leftarrow a_i(k) \frac{\sum_t r_t(k) u^{(a)}_{i,t}}{\sum_t u^{(a)}_{i,t}}, \tag{47}$$

$$u^{(a)}_{i,t} \leftarrow u^{(a)}_{m,t} \frac{\sum_k r_t(k) a_i(k)}{\sum_k a_i(k)}, \tag{48}$$

where

$$\alpha_t(k) = \sum_z u^{(e)}_{z,t} e(k-z), \tag{49}$$

$$\beta_t(k) = \sum_m u^{(h)}_{m,t} h_m(k), \tag{50}$$

$$\gamma_t(k) = \sum_i u^{(a)}_{i,t} a_i(k), \tag{51}$$

$$r_t(k) = \frac{y_t(k)}{\alpha_t(k)\beta_t(k) + \gamma_t(k)}. \tag{52}$$